# Research Development

## AR, k-class, double k-class estimator

Will (Jinwei) Zhang

University of Washington

March 11, 2023

# Table of contents

# Introduction

Problem:

- Casual Models are not prediction optimal under bounded perturbations
- 2 stage least squares (2SLS) suffers small sample bias with weak instruments
- 2SLS has a larger variance structurally
- The relationship between target variables, confounders, and endogenous variables is not certain.

Solution:

- Proposed modified double k-class estimator (DKE)

# Linear structural causal model

Given centered variables $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $H \in \mathbb{R}^r$, $A \in \mathbb{R}^q$, let the distribution of $(X, Y, H, A)$ under $\mathbb{P}_{\text{train}}$ be a solution of the SEM

$$\begin{bmatrix} X \\ Y \\ H \end{bmatrix} = \mathbf{B}^\top \begin{bmatrix} X \\ Y \\ H \end{bmatrix} + \varepsilon + \mathbf{M}^\top A \tag{1.1}$$

where $H \in \mathbb{R}^r$ is unobservable (confounder), $\varepsilon \in \mathbb{R}^{d+1+r}$ is noise, $\mathbf{M} \in \mathbb{R}^{q \times (d+1+r)}$ and $\mathbf{B} \in \mathbb{R}^{(d+1+r)\times(d+1+r)}$ are unknown constant matrices.

Assume $\mathrm{Id} - \mathbf{B}^\top$ is invertible. Then distribution of $(X, Y, H, A)$ is well-defined in terms of $\mathbf{B}, \varepsilon, \mathbf{M}$ and $A$ with unique solution

$$\begin{bmatrix} X \\ Y \\ H \end{bmatrix} = (\mathrm{Id} - \mathbf{B}^\top)^{-1}(\varepsilon + \mathbf{M}^\top A) \tag{1.2}$$

# Linear Estimator

Let $(\mathbf{Y}, \mathbf{X}, \mathbf{H}, \mathbf{A})$ consist of $n$ row-wise random vector $(Y, X, H, A)$
Given the SEM, consider linear estimators

$$\tilde{\beta} = \mathbf{CY},$$

where $\mathbf{C} \in \mathbb{R}^{d \times n}$ is arbitrary. The moment matrix of $\tilde{\beta}$ can be written as

$$\begin{aligned}
&\mathbb{E}[(\tilde{\beta} - \beta_0)(\tilde{\beta} - \beta_0)^\top] \\
=&\mathbb{E}[(\mathbf{C}(\mathbf{A}\gamma_0 + \boldsymbol{\epsilon}_{AY}) - \beta_0)(\mathbf{C}(\mathbf{A}\gamma_0 + \boldsymbol{\epsilon}_{AY} - \beta_0)^\top] \\
=&\sigma^2 \mathbf{C}\mathbf{C}^\top + \mathbf{C}\mathbf{A}\gamma_0\gamma_0^\top(\mathbf{C}\mathbf{A})^\top - \mathbf{C}\mathbf{A}\gamma_0\beta_0^\top - \beta_0\gamma_0^\top(\mathbf{C}\mathbf{A})^\top + \beta_0\beta_0^\top
\end{aligned}$$

as we assume i.i.d. $\epsilon$, and $\mathbf{Y} = \mathbf{A}\gamma_0 + \boldsymbol{\epsilon}_{AY}$.

# Solving the linear equation

We are interested in the moment matrix minimization with respect to $\mathbf{C}$, considering possible measures below:

- Frobenius norm
- Spectral norm (potentially?)
- trace (commonly considered)

$$\mathbf{tr}\left(\sigma^2\mathbf{C}\mathbf{C}^\top + \mathbf{C}\mathbf{A}\gamma_0\gamma_0^\top(\mathbf{C}\mathbf{A})^\top - \mathbf{C}\mathbf{A}\gamma_0\beta_0^\top - \beta_0\gamma_0^\top(\mathbf{C}\mathbf{A})^\top\right)$$

therefore, using FOC,

$$2\sigma^2\mathbf{C} + 2\mathbf{C}(\mathbf{A}\gamma_0\gamma_0^\top\mathbf{A}^\top) - 2\beta_0\gamma_0^\top\mathbf{A}^\top = 0.$$

# Minimum mean square estimator

we find the matrix $\mathbf{C}$ that minimizes moment matrix above:

$$\mathbf{C} = \beta_0 \gamma_0^\top \mathbf{A}^\top (\sigma^2 I + \mathbf{A}\gamma_0\gamma_0^\top \mathbf{A}^\top)^{-1}$$

Then, we derive the MMSE using the SEM:

$$\tilde{\beta} = \beta_0 \gamma_0^\top \mathbf{A}^\top (\sigma^2 I + \mathbf{A}\gamma_0\gamma_0^\top \mathbf{A}^\top)^{-1}\mathbf{Y}$$

Note Woodbury matrix identity gives

$$(\sigma^2 I + \mathbf{A}\gamma_0\gamma_0^\top \mathbf{A}^\top)^{-1} = \frac{1}{\sigma^2}[I - \mathbf{A}\gamma_0(\sigma^2 + \gamma_0^\top \mathbf{A}^\top \mathbf{A}\gamma_0)^{-1}\gamma_0^\top \mathbf{A}^\top]$$

Finally, we can express

$$\tilde{\beta} = \frac{(\mathbf{Y} - \epsilon_{AY})^\top \mathbf{Y}}{\sigma^2 + (\mathbf{Y} - \epsilon_{AY})^\top (\mathbf{Y} - \epsilon_{AY})}\beta_0$$

# Proposal on sample analogue

For sample analogue of $\gamma_0$ and $\sigma^2$:

- Given the SEM and assumptions, we use OLS, so $\hat{\gamma} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{Y}$
- Then, define $P_{\mathbf{A}} = \mathbf{A} \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top$ as the projection onto the d-dimensional range of $\mathbf{A}$.
- we have $\hat{\sigma}^2 = \frac{1}{n-d} \epsilon_{AY}^\top \epsilon_{AY} = \frac{1}{n-d} \mathbf{Y}^\top P_{\mathbf{A}}^\perp \mathbf{Y}$, where $P_{\mathbf{A}}^\perp = I - P_{\mathbf{A}}$.

$\beta_0$ remains unsolved to produce a class of biased estimators.

- Define $\mathbf{Z} = [\mathbf{X}\ \mathbf{A}_*]$, where $\mathbf{A}_* \subset \mathbf{A}$ are the included exogenous variables.
- I propose a k-class estimator as the sample analogue

$$\hat{\beta}_\kappa = \left(\mathbf{Z}^\top \left(I - \kappa P_{\mathbf{A}}^\perp\right) \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \left(I - \kappa P_{\mathbf{A}}^\perp\right) \mathbf{Y},$$

# Proposal on $\tilde{\beta}$

Then, we have the final proposed estimator

$$\beta^* = \frac{(\mathbf{Y} - \boldsymbol{\epsilon}_{AY})^\top \mathbf{Y}}{\hat{\sigma}^2 + (\mathbf{Y} - \boldsymbol{\epsilon}_{AY})^\top (\mathbf{Y} - \boldsymbol{\epsilon}_{AY})} \hat{\beta}_\kappa$$

And alternatively:

$$\beta^* = \left[ 1 - \frac{\boldsymbol{\epsilon}_{AY}^\top \boldsymbol{\epsilon}_{AY} / (n-d)}{\mathbf{Y}^\top \mathbf{Y} - \boldsymbol{\epsilon}_{AY}^\top \boldsymbol{\epsilon}_{AY} [1 - (1/(n-d))]} \right] \hat{\beta}_\kappa$$

# Intuition

- Rothenhäusler *et al.* (2018) present Anchor regression (AR) that exhibits distributional robustness under bounded perturbations. In fact, AR is simply a special case of the k-class estimators, having the same robustness properties given the required assumptions (Jakobsen and Peters, 2021).
- AR as a penalized regression uses 2SLS, and requires under-identification for its theorem to hold. These raise questions about its efficiency, especially under small samples.
- Ullah and Ullah (1978) proposed double k-class estimator that dominates OLS in terms of risk function under certain cases. I consider applying the SEM with endogenous variables to DKE as an alternative estimator to AR. The properties of the application, biases, and risk function expressions are yet to be researched.

. . .

# K-class Estimator (Jakobsen and Peters, 2021)

Given nonsample information about which components of $\beta_0$ and $\gamma_0$ are zero, partition $\mathbf{X} = [\mathbf{X}_*\mathbf{X}_{-*}] \in \mathbb{R}^{n\times(d_1+d_2)}$, $\mathbf{A} = [\mathbf{A}_*\mathbf{A}_{-*}] \in \mathbb{R}^{n\times(q_1+q_2)}$ and $\mathbf{Z} = [\mathbf{Z}_*\mathbf{Z}_{-*}] = [\mathbf{X}_*\mathbf{A}_*\mathbf{X}_{-*}\mathbf{A}_{-*}]$ with $\mathbf{Z} \in \mathbb{R}^{n\times((d_1+q_1)+(d_2+q_2))}$, where $\mathbf{X}_{-*}$ and $\mathbf{A}_{-*}$ correspond to the zero-components of $\beta_0$ and $\gamma_0$. Similarly, $\gamma_0 = (\gamma_{0,*}, \gamma_{0,-*})$,$\beta_0 = (\beta_{0,*}, \beta_{0,-*})$,$\alpha_0 = (\alpha_{0,*}, \alpha_{0,-*}) = (\beta_{0,*}, \gamma_{0,*}, \beta_{0,-*}, \gamma_{0,-*})$. Interested equation:

$$\mathbf{Y} = \mathbf{X}_*\beta_{0,*} + \mathbf{X}_{-*}\beta_{0,-*} + \mathbf{A}_*\gamma_{0,*} + \mathbf{A}_{-*}\gamma_{0,-*} + \tilde{\mathbf{U}}_Y = \mathbf{Z}_*\alpha_{0,*} + \mathbf{U}_Y$$

where $\mathbf{U}_Y = \mathbf{X}_{-*}\gamma_{0,-*} + \mathbf{A}_{-*}\beta_{0,-*} + \mathbf{H}\eta_0 + \varepsilon_Y$. When well-defined, with parameter $\kappa \in \mathbb{R}$ for a simultaneous estimation of $\alpha_{0,*}$:

$$\hat{\alpha}_{\mathrm{K}}^{n}\left(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}\right) = \left(\mathbf{Z}_*^\top\left(I - \kappa P_{\mathbf{A}}^{\perp}\right)\mathbf{Z}_*\right)^{-1}\mathbf{Z}_*^\top\left(I - \kappa P_{\mathbf{A}}^{\perp}\right)\mathbf{Y},$$

where $I - \kappa P_{\mathbf{A}}^{\perp} = I - \kappa\left(I - P_{\mathbf{A}}\right) = (1-\kappa)I + \kappa P_{\mathbf{A}}$.

# Anchor Regression (Rothenhäusler *et al.*, 2018)

K-class estimator with no included exogenous variables, for $\kappa < 1$ coincides with the AR estimator with penalty parameter $\lambda = \kappa/(1-\kappa)$, i.e., $\hat{\beta}^n_{\mathrm{K}}(\kappa) = \hat{\beta}^n_{\mathrm{AR}}\left(\frac{\kappa}{1-\kappa}\right)$, or Equivalently $\hat{\beta}^n_{\mathrm{AR}}(\lambda) = \hat{\beta}^n_{\mathrm{K}}(\lambda/(1+\lambda))$ for any $\lambda > -1$.
The solution to AR empirical minimization problem can be written as

$$\hat{\beta}^n_{\mathrm{AR}}(\lambda) = \left[\mathbf{X}^\top (I + \lambda P_{\mathbf{A}})\, \mathbf{X}\right]^{-1} \mathbf{X}^\top (I + \lambda P_{\mathbf{A}})\, \mathbf{Y},$$

AR aim to possess a interventional robustness instead of inferring causality.

$$\gamma_{\mathrm{AR}}(\lambda) = \underset{\gamma \in \mathbb{R}^d}{\arg\min} \sup_{v \in C(\lambda)} E^{\mathrm{do}(A:=v)} \left[\left(Y - \gamma^\top X\right)^2\right],$$

$$C(\lambda) := \left\{v : \Omega \to \mathbb{R}^q : \mathrm{Cov}(v, \varepsilon) = 0, E\left(vv^\top\right) \preceq (\lambda+1) E\left(AA^\top\right)\right\}.$$

# Double k-Class Estimator (Ullah and Ullah, 1978)

Consider regression model $y = X\beta + u$, and define

$$b = (X'X)^{-1} X'y,$$
$$s^2 = \frac{1}{n}\hat{u}'\hat{u} = \frac{1}{n}y'My,$$

and a class of linear estimators $\beta^* = Ay$. A sample analogue to the MMSE estimator is proposed:

$$\tilde{b} = \left[1 - \frac{\hat{u}'\hat{u}/n}{y'y - \hat{u}'\hat{u}[1-(1/n)]}\right] b$$

A natural generalization of $\tilde{b}$ is the following double $k$-class estimators:

$$\tilde{b}_{k_1,k_2} = \left[1 - \frac{k_1\hat{u}'\hat{u}}{y'y - k_2\hat{u}'\hat{u}}\right] b$$

where $k_1, k_2$ are arbitrary scalars which may be stochastic or non-stochastic.

Let $P_x = X\left(X'X\right)^{-1}X'$ be the orthogonal projection onto the column space of $X$, and $M = I - P_x$

$$\begin{aligned}
\tilde{b}_{k_1,k_2} &= \left[1 - \frac{k_1 \hat{u}'\hat{u}}{y'y - k_2\hat{u}'\hat{u}}\right] b \\
&= \left[1 - \frac{k_1 y'My}{y'y - k_2 y'My}\right] \left(X'X\right)^{-1} X'y \\
&= \left[1 - \frac{k_1 y'(I - P_x)y}{y'((1-k_2)I + k_2 P_x)y}\right] \left(X'X\right)^{-1} X'y \\
&= \left[1 - \frac{y'[k_1(I - P_x)]y}{y'[I - k_2(I - P_x)]y}\right] \left(X'X\right)^{-1} X'y
\end{aligned}$$

# Linear estimator, propose 2 (abandoned)

Let $(\mathbf{Y}, \mathbf{X}, \mathbf{H}, \mathbf{A})$ consist of $n$ row-wise random vector $(Y, X, H, A)$ and consider the single equation of interest

$$\mathbf{Y} = \mathbf{X}\beta_0 + \mathbf{A}\gamma_0 + \mathbf{H}\eta_0 + \boldsymbol{\epsilon}_Y = \mathbf{X}\beta_0 + \mathbf{A}\gamma_0 + \mathbf{U}_Y$$

Given the equation of interest, consider linear estimators

$$\tilde{\beta} = \mathbf{C}\mathbf{Y},$$

where $\mathbf{C} \in \mathbb{R}^{d \times n}$ is arbitrary.

# Linear estimator, continued

The moment matrix of $\tilde{\beta}$ can be written as

$$\begin{aligned}
&\mathbb{E}[(\tilde{\beta} - \beta_0)(\tilde{\beta} - \beta_0)^\top] \\
&= \mathbb{E}[(\mathbf{C}(\mathbf{X}\beta_0 + \mathbf{A}\gamma_0 + \mathbf{U}_Y) - \beta_0)(\mathbf{C}(\mathbf{X}\beta_0 + \mathbf{A}\gamma_0 + \mathbf{U}_Y) - \beta_0)^\top] \\
&= \mathbb{E}[\mathbf{C}\mathbf{U}_Y\mathbf{U}_Y^\top\mathbf{C}^\top + (\mathbf{C}\mathbf{X} - I)\beta_0\beta_0^\top(\mathbf{C}\mathbf{X} - I)^\top + \mathbf{C}\mathbf{A}\gamma_0\gamma_0^\top(\mathbf{C}\mathbf{A})^\top \\
&\quad + (\mathbf{C}\mathbf{X} - I)\beta_0(\mathbf{C}\mathbf{U}_Y + \mathbf{C}\mathbf{A}\gamma_0)^\top + (\mathbf{C}\mathbf{U}_Y + \mathbf{C}\mathbf{A}\gamma_0)\beta_0^\top(\mathbf{C}\mathbf{X} - I)^\top \\
&\quad + \mathbf{C}\mathbf{A}\gamma_0(\mathbf{C}\mathbf{U}_Y)^\top + \mathbf{C}\mathbf{U}_Y\gamma_0^\top(\mathbf{C}\mathbf{A})^\top]
\end{aligned}$$

Does not seem to work

D. Rothenhäusler, N. Meinshausen, P. Bühlmann and J. Peters, *Anchor regression: heterogeneous data meets causality*, 2018, `https://arxiv.org/abs/1801.06229`.

M. E. Jakobsen and J. Peters, *The Econometrics Journal*, 2021, **25**, 404–432.

A. Ullah and S. Ullah, *Econometrica*, 1978, **46**, 705–722.